

Algebraic reputation model RepRank and its application to spambot detection *

G.V. Ovchinnikov [†]
D.A. Kolesnikov [‡]I.V. Oseledets [§]**Abstract**

Due to popularity surge social networks became lucrative targets for spammers and guerilla marketers, who are trying to game ranking systems and broadcast their messages at little to none cost. Ranking systems, for example Twitter's Trends, can be gamed by scripted users also called bots, who are automatically or semi-automatically twitting essentially the same message. Judging by the prices and abundance of supply from PR firms this is an easy to implement and widely used tactic, at least in Russian blogosphere. Aggregative analysis of social networks should at best mark those messages as spam or at least correctly downplay their importance as they represent opinions only of a few, if dedicated, users. Hence bot detection plays a crucial role in social network mining and analysis.

In this paper we propose technique called RepRank which could be viewed as Markov chain based model for reputation propagation on graphs utilizing simultaneous trust and anti-trust propagation and provide effective numerical approach for its computation.

Comparison with another models such as TrustRank and some of its modifications on sample of 320000 Russian speaking Twitter users is presented. The dataset is presented as well.

Keywords. Antispam, social graph, Markov chain, trustrank.

1 Introduction

While concepts of 'good' and 'bad' are quite complex and may be considered subjective, applied to spam filtration problems they are mapped to 'signal' and 'noise' correspondingly and became objective enough to build algorithms upon.

The following assumption, proposed in [1] was successfully used in graph based antispam algorithms:

Assumption 1 (Approximate isolation of the good set)

Good graph vertices rarely link to the bad ones.

*This work was supported by Russian Science Foundation Grant 14-11-00659

[†]Skolkovo Institute of Science and Technology, Skolkovo 143025, Russia and Institute of Design Problems in Microelectronics, Zelenograd Sovetskaya 3, Moscow, 124365 (e-mail: ovgeorge@yandex.ru)

[‡]Skolkovo Institute of Science and Technology, Skolkovo 143025, Russia (e-mail: d4kolesnikov@yandex.ru)

[§]Skolkovo Institute of Science and Technology, Skolkovo 143025, Russia and Institute of Numerical Mathematics, Russian Academy of Sciences, ul. Gubkina 8, Moscow, 119333 Russia. (e-mail: ivan.oseledets@gmail.com)

This assumption leads to trust propagation scheme proposed in [1]: pages linked from good ones are almost good, the pages linked from those are slightly worse and so on. There is a similar approach to mistrust propagation [2]: pages linking to bad pages are almost bad, the pages linking to those pages may be slightly better and so on. The main difference being trust propagates forward (by graph edges direction) and mistrust propagates backward.

TrustRank and other similar models for signal propagation on graphs can be viewed in random walker framework with random walkers carrying signal. In case of TrustRank it is trust charge equal to current vertex trust value. For some vertices marked by external verification process (also called oracle function) walkers with a priory set probability α take charge equal to 1 and with probability $1 - \alpha$ take charge equal to current charge of the vertex. Stationary distribution of such process with initial distribution vector d with $d_i = 1$ if i -th vertex marked by the oracle as a good one and $d_i = 0$ otherwise is gives us TrustRank scores for whole graph:

$$(1.1) \quad t = \alpha F t + (1 - \alpha) d,$$

here F is forward transition matrix which is column normalized adjacency matrix.

The logical continuation of TrustRank and anti-TrustRank is the combination of both models. Major advantage of combined trust and mistrust propagation approach is the ability to use both positive and negative signals.

[4] uses this approach by penalizing trust and distrust propagation from not trustworthy and trustworthy vertices correspondingly. [3] is analogous to [4], but stated in probabilistic framework.

In this paper we further pursue the idea of simultaneous trust and mistrust propagation. In contrast with above-mentioned papers we combine trust and mistrust providing unified reputation score, called RepRank.

2 Proposed model

To each graph vertex we attach a trust score which is negative for bad vertices and positive for a good ones.

Denote by t_+ and t_- vectors obtained by zeroing negative and positive components in vector t correspondingly. Then, we search trust distribution t to satisfy

$$(2.2) \quad t = \alpha_1 F t_+ + \alpha_2 B t_- + \alpha_3 d,$$

where B is backward transition matrix which is column normalized transposed adjacency matrix. The solution of (2.2) we will call RepRank. It exists, unique and continuously dependent on the initial distribution d (see the Theorem 2.1 for more details). While usefulness of existence and uniqueness of the solution to the equation (2.2) solution is beyond doubt we want to point out, that continuous dependence on oracle provided labeling d is very nice as well. It protects RepRank from sudden "everything you knew is wrong" changes caused by addition of small portion of new data, mistakes and typos (a manual labeling is very tedious and error prone process). This follows natural intuition that once one have an idea of everyone's reputation a small changes should only correct, not shake the foundations of his worldview.

THEOREM 2.1. *Let F, B be $n \times n$ stochastic matrices from (2.2), $0 < \alpha_1, \alpha_2, \alpha_3 < 1$, \mathcal{P}_+ be an operator which replaces all negative component of the $n \times 1$ vector by zeroes and \mathcal{P}_- be operator which replaces all positive components of the $n \times 1$ vector by zeroes. Then mapping \mathcal{R} an $n \times 1$ vector d to the $n \times 1$ vector t that the solution of the equation*

$$(2.3) \quad t = \alpha_1 F \mathcal{P}_+(t) + \alpha_2 B \mathcal{P}_-(t) + \alpha_3 d,$$

has the next properties:

1. $R(d)$ exists for any vector d from \mathbf{R}^n .
2. $R(d)$ is bijection mapping \mathbf{R}^n to \mathbf{R}^n .
3. $R(d)$ is Lipschitz continuous mapping.

Proof:

One can use the following iterative process to find t :

$$\begin{aligned} t^{(k+1)} &= I(t^{(k)}), \\ I(t) &= \alpha_1 F \mathcal{P}_+(t) + \alpha_2 B \mathcal{P}_-(t) + \alpha_3 d, \end{aligned}$$

with any initial vector t_0 . The mapping $I(t)$ is contrac-

tive on the metric space $(\mathbf{R}^n, \|\cdot\|_1)$:

$$\begin{aligned} \|I(t_1) - I(t_2)\|_1 &= \\ &= \|\alpha_1 F(\mathcal{P}_+(t_1) - \mathcal{P}_+(t_2)) + \\ &\quad + \alpha_2 B(\mathcal{P}_-(t_1) - \mathcal{P}_-(t_2))\|_1 \leq \\ &\leq \alpha_1 \|F\|_1 \|\mathcal{P}_+(t_1) - \mathcal{P}_+(t_2)\|_1 + \\ &\quad \alpha_2 \|B\|_1 \|\mathcal{P}_-(t_1) - \mathcal{P}_-(t_2)\|_1 \leq \\ &\leq \alpha_1 \|\mathcal{P}_+(t_1) - \mathcal{P}_+(t_2)\|_1 + \\ &\quad \alpha_2 \|\mathcal{P}_-(t_1) - \mathcal{P}_-(t_2)\|_1 \leq \\ &\leq \max(\alpha_1, \alpha_2) \|t_1 - t_2\|_1, \end{aligned}$$

where we used that $\|F\|_1 = \|B\|_1 = 1$ and

$$\|t_1 - t_2\|_1 = \|\mathcal{P}_+(t_1) - \mathcal{P}_+(t_2)\|_1 + \|\mathcal{P}_-(t_1) - \mathcal{P}_-(t_2)\|_1.$$

So $I(t)$ is contractive mapping with coefficient less or equal to $\max(\alpha_1, \alpha_2) < 1$ and Banach fixed-point theorem guarantees the existence and uniqueness of fixed point for it. Let us denote that fixed point by $R(d)$ and notice that it is the solution for equation (2.2).

Also $R(d)$ is Lipschitz continuous mapping with coefficient

$$\frac{\alpha_3}{1 - \max(\alpha_1, \alpha_2)}.$$

It can be proven in the following way:

$$\begin{aligned} R(d_1) - R(d_2) &= \\ &= \alpha_1 F(\mathcal{P}_+(R(d_1)) - \mathcal{P}_+(R(d_2))) + \\ &\quad + \alpha_2 B(\mathcal{P}_-(R(d_1)) - \mathcal{P}_-(R(d_2))) + \alpha_3(d_1 - d_2) \\ &\quad \|R(d_1) - R(d_2)\|_1 \leq \\ &\leq \alpha_1 \|F\|_1 \|\mathcal{P}_+(R(d_1)) - \mathcal{P}_+(R(d_2))\|_1 + \\ &\quad + \alpha_2 \|B\|_1 \|\mathcal{P}_-(R(d_1)) - \mathcal{P}_-(R(d_2))\|_1 + \\ &\quad + \alpha_3 \|d_1 - d_2\|_1 \leq \\ &\leq \max(\alpha_1, \alpha_2) \|R(d_1) - R(d_2)\|_1 + \alpha_3 \|d_1 - d_2\|_1, \end{aligned}$$

therefore

$$\|R(d_1) - R(d_2)\|_1 \leq \frac{\alpha_3}{1 - \max(\alpha_1, \alpha_2)} \|d_1 - d_2\|_1.$$

The mapping $R(d)$ is injection because equality $R(d_1) = R(d_2)$ causes $d_1 = d_2$ in equation (2.2). It is also a surjection because for any $t \in \mathbf{R}^n$ exists d_t such that

$$d_t = \frac{1}{\alpha_3} (t - \alpha_1 F \mathcal{P}_+(t) - \alpha_2 B \mathcal{P}_-(t)),$$

that equality $R(d_t) = t$ holds.

3 Experimental evaluation

For our experiments we recursively crawled twitter using as seeds Russian speaking users we found in Twitter's Streaming API. For each user we downloaded all people

| Algorithm | Accuracy |
|----------------|----------|
| RepRank | 0.8833 |
| TrustRank | 0.851 |
| anti-TrustRank | 0.8636 |

Table 1: Experimental results

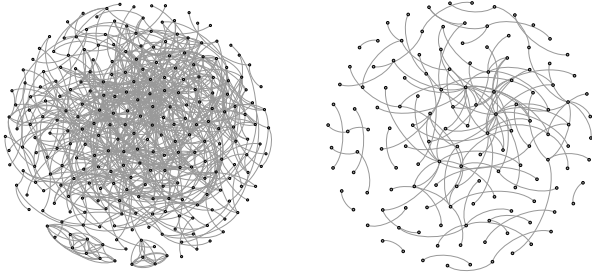


Figure 1: Connections between 300 vertices with highest in-degree (left) and 300 most reputable vertexes (right). Vertices not having connections within this group are omitted.

he follows, 'friends' in Twitter terminology. This friends graph has 326130 vertices and 2713369 nodes. We manually labeled 3124 users as spammers or a good ones.¹

We did a cross-validation with random subsampling splitting data set in two halves to test our algorithm against other single-score trust propagation algorithms, namely TrustRank and anti-TrustRank. The parameters for each algorithm were chosen to maximize accuracy.

The results provided in the Table 1. As can be seen our algorithm outperforms both TrustRank and anti-TrustRank.

It is interesting to note, that Russian-speaking part of Twitter is dominated by bots and spammers. According to our algorithm, out of 326130 accounts only 59691 (around 18%) are managed by humans. Among those only 375 (around 0.1%) correspond to high-reputation, profiles of prominent public figures, government officials and organizations, reputable press and so on (see reputation distribution on Figure 2). Spammers are more active at following each other than a normal people (see Figure 1). A serious anti-spam effort is required if one wants to make use of Twitter's

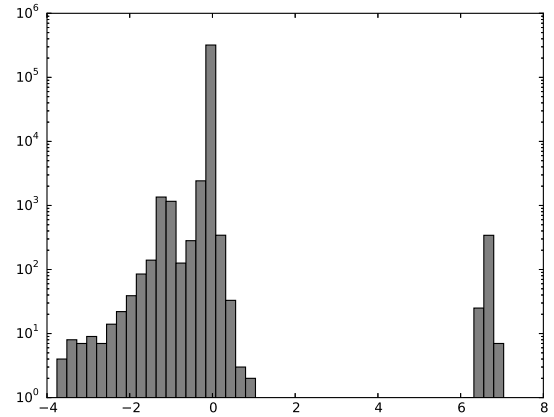


Figure 2: Distribution of reputation. RepRank score on the x-axis and logarithm of number of accounts on the y-axis. The gap is due to separation between masses and a handful of celebrities who are following each other.

data, otherwise he will analyze noise, or worse, some botmaster's political views.

4 Conclusion and further work

The contribution of this paper is twofold. First, we proposed new reputation propagation algorithm which allows to use both bad and good vertices in its starting set and outperforms analogues. Second, we gathered a sample Russian Twitter social graph with manual labeling of good and bad seeds.

Different normalization strategies along with regularizations (pagerank-style teleportations) could be used to further improve performance of proposed method.

References

- [1] Zoltán Gyöngyi, Hector Garcia-Molina, and Jan Pedersen. Combating web spam with trustrank. In *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30, VLDB '04*, pages 576–587. VLDB Endowment, 2004.
- [2] Vijay Krishnan. Web spam detection with anti-trust rank. In *In AIRWEB*, pages 37–40, 2006.
- [3] Xinyue Liu, You Wang, Shaoping Zhu, and Hongfei Lin. Combating web spam through trust-distrust propagation with confidence. *Pattern Recogn. Lett.*, 34(13):1462–1469, October 2013.
- [4] Xianchao Zhang, You Wang, Nan Mou, and Wenxin Liang. Propagating both trust and distrust with

¹The dataset can be obtained from <https://bitbucket.org/ovchinnikov/rutwitterdataset>

target differentiation for combating link-based web spam. *ACM Trans. Web*, 8(3):15:1–15:33, July 2014.